

This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

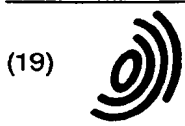
Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning documents *will not* correct images,  
please do not report the images to the  
Image Problem Mailbox.**

**THIS PAGE BLANK (USPTO)**



Europäisches Patentamt  
European Patent Office  
Office européen des brevets



(11) EP 0 701 208 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:  
13.03.1996 Bulletin 1996/11

(51) Int. Cl.<sup>6</sup>: G06F 11/10, G11B 20/18

(21) Application number: 95114142.3

(22) Date of filing: 08.09.1995

(84) Designated Contracting States:  
DE FR GB

(30) Priority: 09.09.1994 JP 215590/94

(71) Applicant: HITACHI, LTD.  
Chiyoda-ku, Tokyo 100 (JP)

(72) Inventors:  
• Kojima, Akira  
Odawara-shi, Kanagawa-ken 256 (JP)

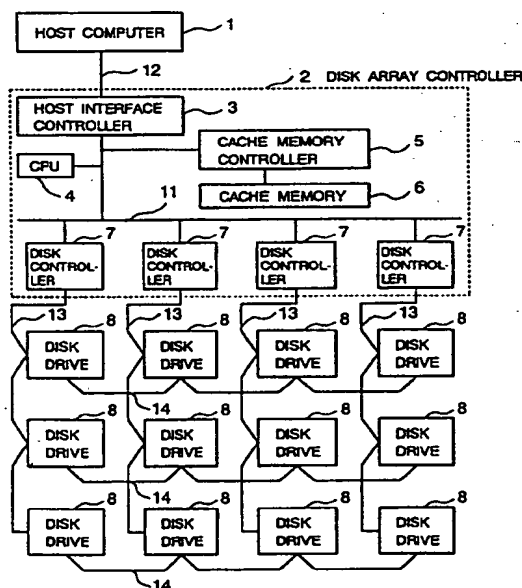
• Ogino, Akihito  
Kanagawa-ken 259-01 (JP)  
• Isono, Soichi  
Chigasaki-shi, Kanagawa-ken 253 (JP)

(74) Representative: Strehl Schübel-Hopf Groening &  
Partner  
Maximilianstrasse 54  
D-80538 München (DE)

(54) Disk array subsystem and data generation method therefor

(57) A disk drive has two channels of upper interfaces, and one channel is connected to a disk array controller, and the other channel is connected between a plurality of disk drives. A data disk drive reads the old data on the recording medium, calculates the exclusive OR of the old data and the corresponding data from the disk array controller, and transfers it to a parity disk as pseudoparity data from the other channel. The parity disk drive reads the old parity data on the recording medium, calculates the exclusive OR of the old parity data and the pseudo-parity data, and writes it as new parity data.

FIG. 1



EP 0 701 208 A2

## Description

### BACKGROUND OF THE INVENTION

#### Field of the Invention

The present invention relates to a disk array subsystem for connecting and operating a plurality of disk drives in parallel and a data generation method therefor and particularly to a disk array subsystem for realizing a reduction in a write penalty in RAID4 (redundant array of inexpensive disks) and RAID5.

### DESCRIPTION OF THE PRIOR ART

A disk array subsystem is a magnetic disk drive for connecting and operating a plurality of inexpensive small disk drives in parallel so as to realize performance equivalent to that of a SLED (single Large expensive disk). A general constitution of disk array is shown in Fig. 7. This disk array comprises a disk array controller 2 which is connected to a host computer 1 via a host interface 12 and a plurality of disk drives 18 which are connected to the disk array controller 2 and operate in parallel. The disk array controller 2 comprises a host interface controller 3 for storing a read or write instruction from the host computer 1 once, a CPU 4 for controlling the operation of the disk array controller 2, a cache memory 6 for storing data transferred between the host computer 1 and the disk drives 18, a cache memory controller 5 for controlling the cache memory, and a disk controller 7 for controlling data transfer between the disk array controller 2 and the disk drives 18.

When it is confirmed by the cache memory controller 5 for reading that requested data exists in the cache memory 6, the data is transferred from the cache memory 6 to the host computer 1 via the host interface 12. When the requested data does not exist in the cache memory 6, the CPU 4 stores the data in the cache memory 6 from a disk drive 18 storing the data via the disk controller 7 and the cache memory controller 5. The cache memory controller 5 transfers the data to the host computer 1 after the storing ends or in parallel with the storing.

Write data transferred from the host computer 1 for writing is stored in the cache memory 6 by the cache memory controller 5 via the host interface 12 and the host interface controller 3. The cache memory controller 5 writes the write data into the disk drive 18 designated by the CPU 4 via the disk controller 7 after the storing ends or in parallel with the storing.

To maintain reliability, the disk array subsystem generates parity on data stored on a plurality of data disks and stores it on a parity disk. In RAID4, the parity disk is fixed in a special disk drive. In RAID5, in consideration of that the performance reduces because access is made concentratedly on the parity disk, the parity is distributed evenly to all the disk drives for each data.

USA Patent 5191584 discloses a data updating method in a disk array subsystem of RAID4 or RAID5. According to this data updating method, when a disk array controller accesses one data disk for each processing data, there is no need to access all the data disks even for writing data. The disk array controller calculates new parity data by the exclusive OR of the old data of the data disk for writing, the old parity data of the parity disk, and new data transferred from the host computer and updates the parity disk according to the new parity data. Therefore, another process can be executed for disks other than the data disk for writing and the parity disk. Particularly in RAID5, no parity disk is specified, so that the write process can be executed at the same time.

A problem caused by this method is that the five processes indicated below are generated because the parity disk is updated for writing data as shown in Fig. 8 and the processing capacity lowers.

- 1) Reading the old data from the data disk at the corresponding address
- 2) Reading the old parity data from the parity disk at the corresponding address
- 3) Writing new data on the data disk at the corresponding address
- 4) Calculating the exclusive OR of the new data, old data, and old parity data and obtaining new parity data
- 5) Writing the new parity data on the parity disk

The four processes except 4) among the aforementioned five processes are accompanied by an access process to disk drives and it causes a reduction performance of the disk array subsystem. This reduction performance due to an increase in the access process to disk drives for updating of the parity disk which is generated for writing data is called a write penalty.

### SUMMARY OF THE INVENTION

An object of the present invention is to provide a disk array subsystem for preventing the performance from lowering even if the access process to a recording medium is increased by a write penalty for writing data and a method therefor.

To accomplish the above object, in the disk array subsystem of the present invention comprising a plurality of magnetic disk storage means, at least one magnetic disk storage means among the above magnetic disk storage means acquires data (new data) to be written into one magnetic disk storage means among the above plurality of magnetic disk storage means, reads the data (old data) in the own magnetic disk storage means which corresponds to the new data, calculates the exclusive OR of the new data and the old data, generates the calculated data, and writes the calculated data in the corresponding address in the own magnetic disk storage means or transfers it to the upper apparatus or another magnetic disk storage means.

A disk drive constituting the disk array subsystem of the present invention has at least two interfaces. The first interface connects the disk controller and a disk drive and the second interface connects a plurality of disk drives having parity in common to each other.

New data transferred from the host computer is transferred to a disk drive by the disk array controller. The disk drive receives the new data by the first interface, reads the old data on the recording medium at the same time, calculates the exclusive OR of the new data and the old data, transfers the result from the second interface to the parity disk as pseudo-parity data, and writes the new data on the recording medium. The disk drive in which the parity data of the new data is stored receives the pseudo-parity data from the second interface, reads the old parity data on the recording medium at the same time, calculates the exclusive OR of the pseudo-parity data and the old parity data, and writes the result on the recording medium as new parity data.

The present invention having the aforementioned constitution has the function and operation indicated below.

According to the present invention, a disk drive executes a process for reading old data and old parity data for updating parity data of the disk array subsystem and a process for calculating and writing new parity data, so that the burden imposed on the disk array controller is lightened. Pseudo-parity data is transferred not via the disk array controller, so that a reduction in the processing capacity due to concentrated access to a disk drive from the disk controller can be prevented and the performance for writing data is improved.

The foregoing and other objects, advantages, manner of operation and novel features of the present invention will be understood from the following detailed description when read in connection with the accompanying drawings.

#### BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of a disk array subsystem in an embodiment of the present invention,

Fig. 2 is a detailed block diagram of a disk drive in an embodiment of the present invention,

Fig. 3 is a block diagram of RAID4 in an embodiment of the present invention,

Fig. 4 is a time chart of the operation of RAID4 for writing in an embodiment of the present invention,

Fig. 5 is a block diagram of RAID5 in an embodiment of the present invention,

Fig. 6 is a time chart of the operation of RAID5 for writing in an embodiment of the present invention,

Fig. 7 is a block diagram of a conventional disk array subsystem, and

Fig. 8 is a time chart of data transfer of RAID5 for writing in a conventional example.

#### DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Fig. 1 shows an embodiment of the present invention. In this embodiment, a parity group comprises  $n$  data disk drives and one parity disk drive. Each disk drive 8 has two interfaces. One of them is an interface 13 connected to the host computer via a disk controller 7. The other is an interface 14 for connecting disk drives constituting the parity group to each other. In this embodiment, the interface 13 is referred to as a disk interface and the interface 14 is referred to as a parity interface.

A disk drive 8 which receives a write instruction calculates the exclusive OR of new data transferred by the disk interface 13 and the old data on the disk medium and transfers the calculation result to the parity disk drive via the second interface. The parity disk drive stores the exclusive OR of the calculation result and the old parity data stored in the disk drive as new parity data.

Fig. 2 shows a detailed constitution of a disk drive 8. The disk drive 8 comprises a buffer memory 21 for storing data transferred from the disk interface 13 or the parity interface 14, a buffer memory 22 for storing data read from a recording medium, an arithmetic unit 23 for calculating the exclusive OR of output of the buffer memory 21 and output of the buffer memory 22, a switching unit 25 for switching the connection destination of the parity interface 14 or the disk interface 13, a switching unit 24 for switching the connection destination of the disk, interface controllers 28a and 28b between the host interface (HOST IF) and the parity interface 14 (PARITY IF), command interpreters 29a and 29b, and a data speed adjustment FIFO unit 25. The buffer memories 21 and 22 have data speed adjustment FIFO units 30 and 31 for input and output. The disk interface has a data speed adjustment FIFO unit 32 for input and output.

When the disk drive is a data disk, the buffer memory 21 stores the exclusive OR of the write data (new data) transferred from the disk interface 13 and the old data at the corresponding address which is read from the disk and when the disk drive is a parity disk, the buffer memory 21 stores the exclusive OR of the pseudo-parity data transferred from the parity interface 14 and the old parity data at the corresponding address which is read from the disk.

The switching units 24 and 25 select EOR23 for a write penalty countermeasure of the RAID of the present invention. For executing normal read/write, they select OLD/NEW.

The buffer memory 22 stores the old data or old parity data of the recording medium 26. When the disk drive is a data disk, the exclusive OR of the new data and old data is stored in the buffer memory 21, selected by the switching unit 25, and outputted to the parity interface 14. When the disk drive is a parity disk, the exclusive OR of the pseudoparity data and old parity data is stored on the recording medium 26 as new parity data from the FIFO unit 32.

Fig. 3 shows an embodiment of the present invention and the constitution in the case of RAID4. In RAID4, the parity disks are fixed. In this embodiment, the disk drives  $81(n+1)$  to  $8-m(n+1)$  are parity disks.

Fig. 4 shows detailed processing of this embodiment. The disk drive 8-11 which receives write data to address #0 from the disk controller 7-1 via the interface 13-1:

- 1) reports the end to the disk controller 7-1 at the point of time when all the data is received in the buffer memory 21 in the disk drive, 10
- 2) reads the old data at address #0 on the recording medium, (this process may be read in the buffer memory B beforehand or may be executed in parallel with writing new data into the buffer), 15
- 3) calculates the exclusive OR of new data and the old data,
- 4) assumes the calculation result as pseudo-parity and transfers it to the disk drive for storing parity data at the corresponding address via the interface 14-1, and 20
- 5) writes the new data at address #0 on the recording medium 26. 25

The parity disk drive 8-1(n+1) which receives the pseudoparity from the data disk drive 8-11 via the interface 14-1:

- 1) reports the end to the data disk drive 8-11 at the point of time when all the data is received in the buffer memory 21 in the disk drive, 30
- 2) reads the old parity data at address #0nP on the recording medium 26, (this process may be read in the buffer memory 22 beforehand or may be executed in parallel with writing the pseudo-parity data into the buffer), 35
- 3) calculates the exclusive OR of the pseudo-parity data and the old parity data, and
- 4) assumes the calculation result as new parity data and writes it at address #0nP on the recording medium 26. 40

According to the aforementioned processes, the disk drive sends an end report when the data from the upper apparatus is stored in the buffer memory of the disk drive, so that the processing time for writing of data can be shortened viewed from the upper apparatus.

Fig. 5 shows another embodiment of the present invention and the constitution in the case of RAID5. In RAID5, the parity disks are not fixed and the data disks of  $(n+1)s$  rotate for each data processing. The other operations are the same as those in the case of RAID4. It is necessary for a data disk drive to recognize the corresponding parity disk among the disk drives so as to transfer pseudo-parity. As a means therefor, there is a method in which a disk controller instructs it in a write instruction or there is a method in which it is calculated from the remainder when the address is divided by  $(n+1)$

and the ID number of the disk drive itself because the parity data is shifted sequentially as shown in this embodiment.

Fig. 6 shows detailed processing in the case of RAID5. The disk drive 8-11 which receives write data to address #0 from the disk controller 7-1 via the interface 13-1:

- 1) reports the end to the disk controller 7-1 at the point of time when all the data is received in the buffer memory 21 in the disk drive,
- 2) reads the old data at address #0 on the recording medium 26, (this process may be read at another address of the buffer memory B beforehand or may be executed in parallel with writing new data into the buffer),
- 3) calculates the exclusive OR of new data and the old data,
- 4) assumes the calculation result as pseudo-parity and transfers it to the disk drive for storing parity data at the corresponding address via the interface 14-1, and
- 5) writes the new data at address #0 on the disk medium.

The parity disk drive 8-1(n+1) which receives the pseudoparity from the data disk drive 8-11 via the interface 14-1:

- 1) reports the end to the data disk drive at the point of time when all the data is received in the buffer memory 21 in the disk drive,
- 2) reads the old parity data at address #0P on the recording medium 26, (this process may be read in the buffer memory 22 beforehand or may be executed in parallel with writing the pseudo-parity data into the buffer),
- 3) calculates the exclusive OR of the pseudo-parity data and the old parity data, and
- 4) assumes the calculation result as new parity data and writes it at address #0nP on the recording medium 26.

Since the parity disks are not fixed in RAID5, when the write process is being executed for the disk drive 8-11 in which a request address (for example, #0) of the host computer exists and the disk drive 8-1(n+1) in which the parity data exists, access to address  $\#(2n+1)$  of another disk drive 8-13 and the disk drive 8-12 in which the parity data exists is possible. Therefore, when the system is designed so that a parity update process on the parity interface and a data write/read process on the controller interface can be executed by a plurality of disk drives in parallel, the system performance can be improved further.

According to the disk array subsystem of the present invention, updating of parity data for writing data is executed via the interface between the disk drives after a disk drive sends an end report to the disk controller, so

that the delay time due to a write penalty can be shortened viewed from the host computer.

## Claims

1. A disk array subsystem having a plurality of disk storage means comprising:
  - at least one disk storage means among said disk storage means comprising;
    - means for writing new data into one disk storage means among said plurality of disk storage means comprising:
      - means for reading the old data in the own disk storage
        - means which corresponds to said new data;
          - means for calculating the exclusive OR of said new data and said old data and generating calculated data;
            - means for transferring said calculated data to an upper apparatus or another disk recording means as pseudo-parity data when said new data is write data from said upper processing apparatus; and
              - means working when said new data is pseudo-parity data transferred from another disk storage means, for calculating the exclusive OR of said pseudo-parity data and said old data and writing said calculated data into said own disk storage means.
  - 2. A disk array subsystem according to Claim 1, wherein at least said one disk storage means which writes said new data further comprising:
    - a first interface to be connected to an upper processing apparatus; and
      - a second interface to be connected to at least said own disk storage means.
  - 3. A disk array subsystem having a plurality of magnetic disk storage means comprising:
    - a plurality of magnetic disk storage means constituting a parity group among said magnetic disk storage means;
      - said plurality of magnetic disk storage means constituting a parity group comprising;
        - means for writing new data into one magnetic disk storage means among said plurality of magnetic disk storage means comprising:
          - means for reading the old data in said own magnetic disk storage means which corresponds to said new data;
            - means for calculating the exclusive OR of said new data and said old data and generating calculated data;
              - means for transferring said calculated data to an upper apparatus or another disk recording means as pseudo-parity data when said new data is write data from said upper processing apparatus; and
                - means working when said new data is pseudo-parity data transferred from another mag-
    - 4. A disk array subsystem according to Claim 3, wherein at least said one magnetic disk storage means which acquires said new data comprising:
      - a first interface to be connected to an upper processing apparatus; and
        - a second interface to be connected to at least said plurality of magnetic disk storage means constituting a parity group.
    - 5. A disk array subsystem according to Claim 3, wherein said one magnetic disk storage means which generates said pseudoparity data further comprising:
      - a first interface to be connected to an upper processing apparatus;
        - a second interface for connecting said storage means constituting said parity group to each other;
          - a first buffer memory for storing data read from said own magnetic disk storage means;
            - arithmetic means for calculating the exclusive OR of write data transferred from said first and second interfaces and output of said first buffer memory; and
              - a second buffer memory for storing said calculation result.
      - 6. A disk array subsystem according to Claim 5, further comprising:
        - when said write data is new data which is transferred from an upper processing apparatus via said first interface,
          - means for storing the old data updated by said new data in said first buffer memory;
            - means for calculating the exclusive OR of said new data and said old data by said arithmetic means and transferring said calculation result to said upper processing apparatus as parity data via said first interface or to another magnetic disk storage means constituting said parity group via said second interface;
              - when said write data is pseudo-parity data which is transferred from said upper processing apparatus via said second interface,
                - means for storing the old parity data updated by said pseudo-parity data in said first buffer memory;
                  - means for calculating the exclusive OR of said pseudoparity data and said old data by said arithmetic means and storing it in said second buffer memory; and
                    - means for recording said calculation result on a recording medium as parity data.

7. A parity data generation method by at least one magnetic disk storage means among a plurality of magnetic disk storage means in a disk array subsystem having said plurality of magnetic disk storage means comprising:

a step of acquiring data (new data) to be written into one magnetic disk storage means among said plurality of magnetic disk storage means;

a step of reading the data (old data) in the own magnetic disk storage means which corresponds to said new data;

a step of calculating the exclusive OR of said new data and said old data and generating calculated data; and

a step of writing said calculated data into the corresponding address in the own magnetic disk storage means.

magnetic disk storage means among a plurality of magnetic disk storage means in a disk array subsystem having said plurality of magnetic disk storage means comprising:

a step of acquiring data (new data) to be written into one magnetic disk storage means among said plurality of magnetic disk storage means;

a step of reading the data (old data) in the own magnetic disk storage means which corresponds to said new data;

a step of calculating the exclusive OR of said new data and said old data and generating calculated data; and

a step of transferring said calculated data to the upper apparatus or another magnetic disk storage means.

5

10

15

20

25

30

35

40

45

50

55



**FIG. 1**

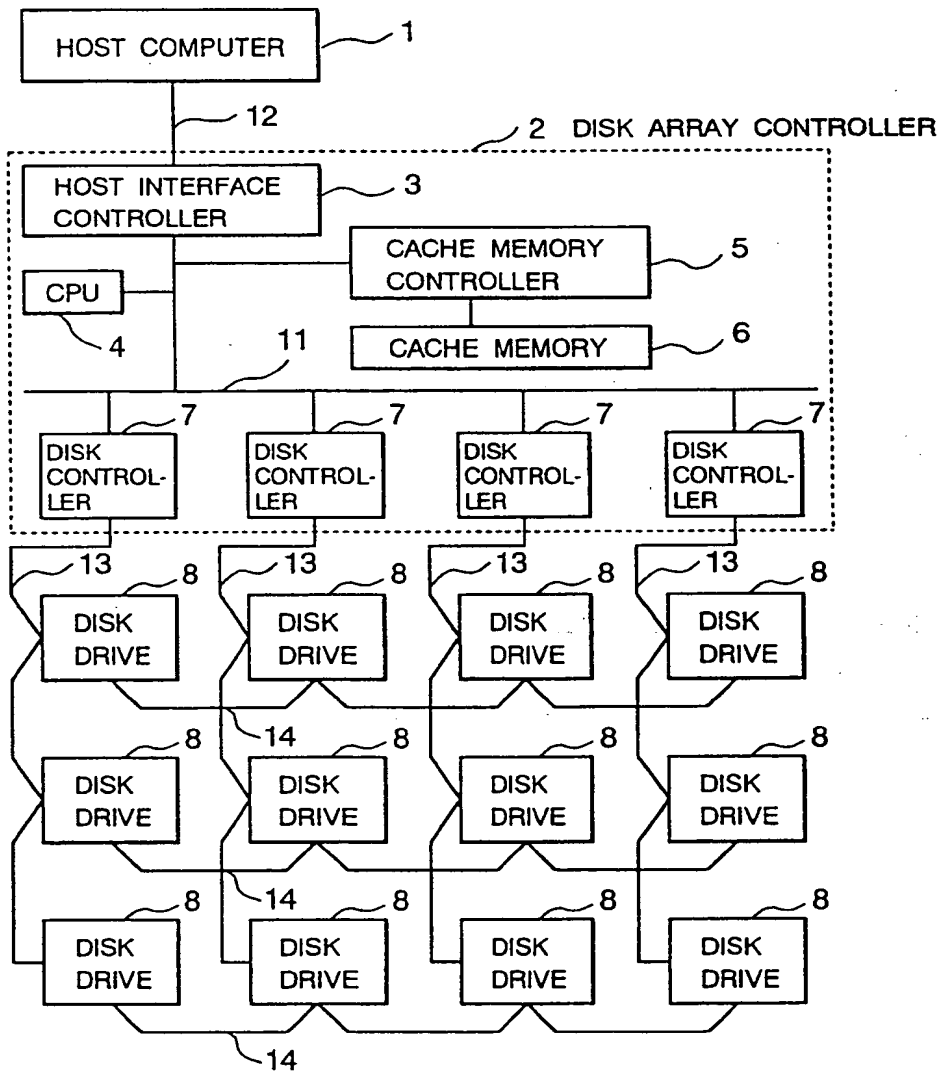


FIG. 2

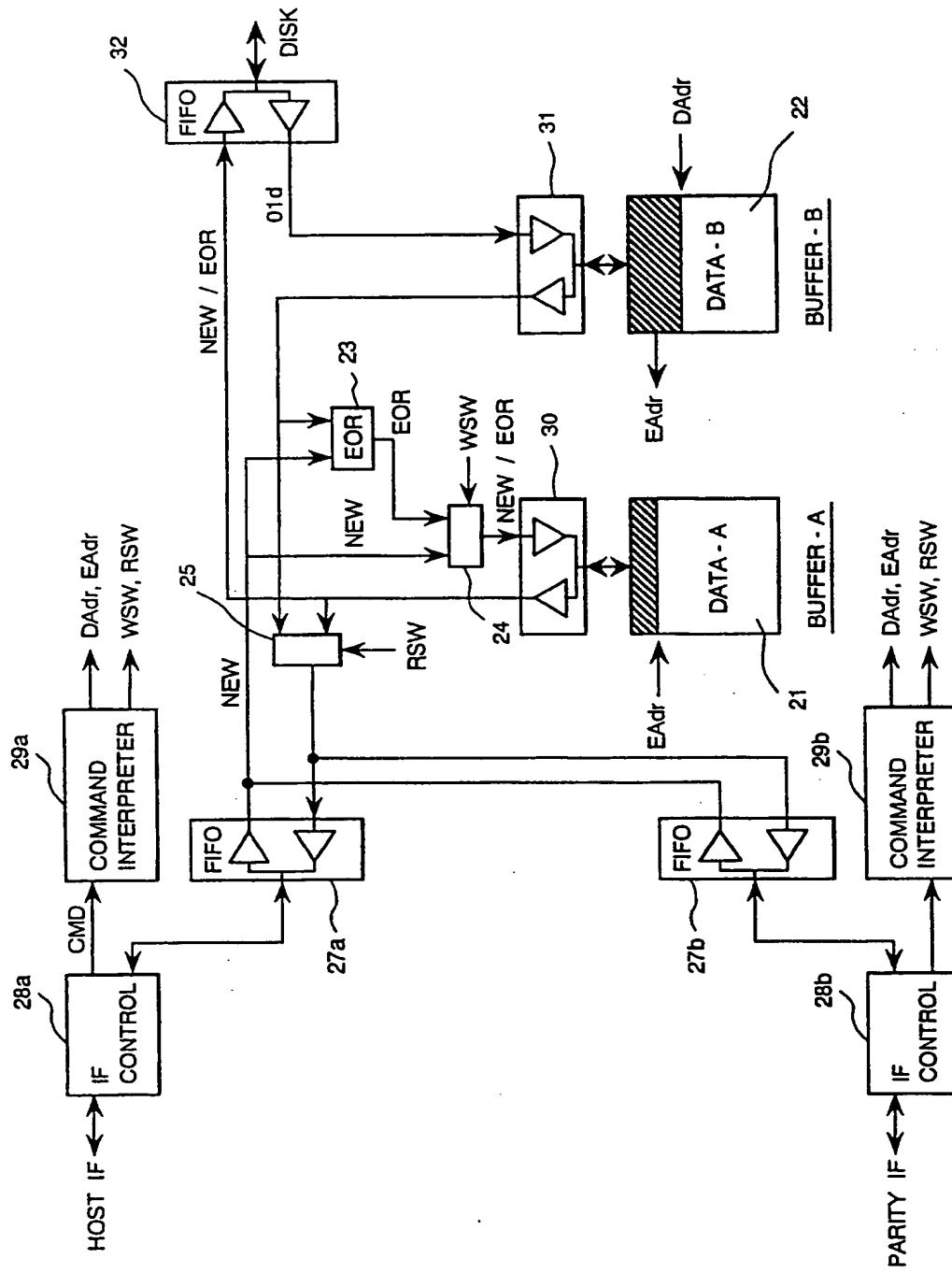


FIG. 3

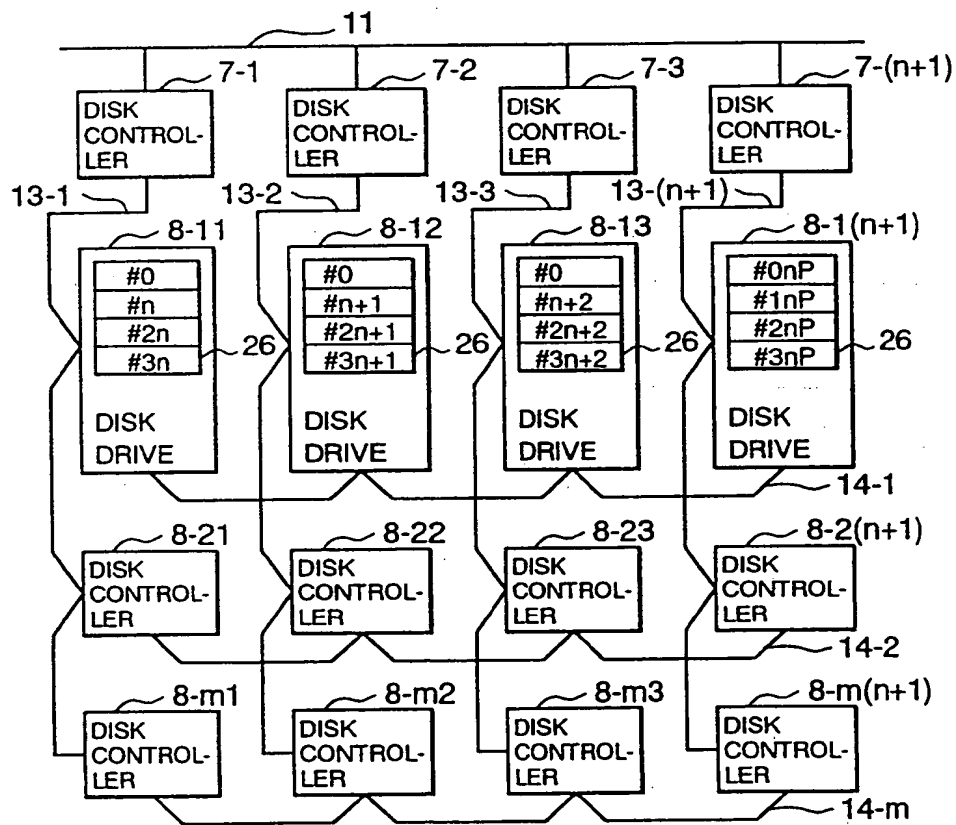


FIG. 4

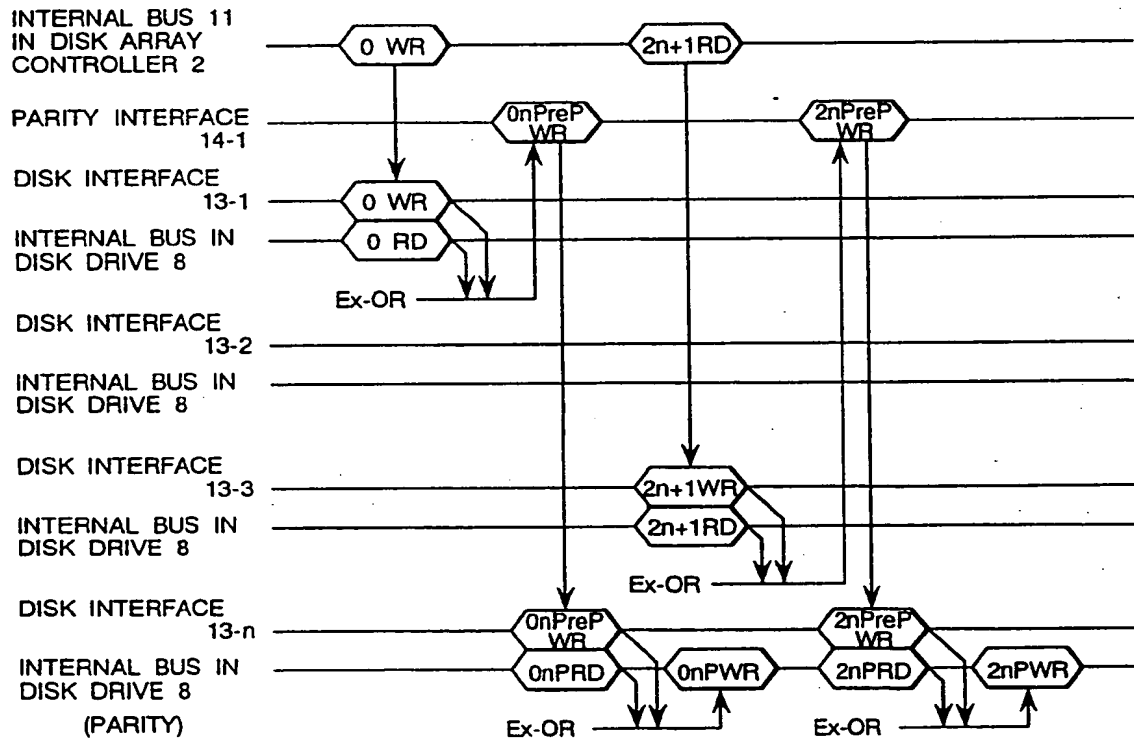


FIG. 5

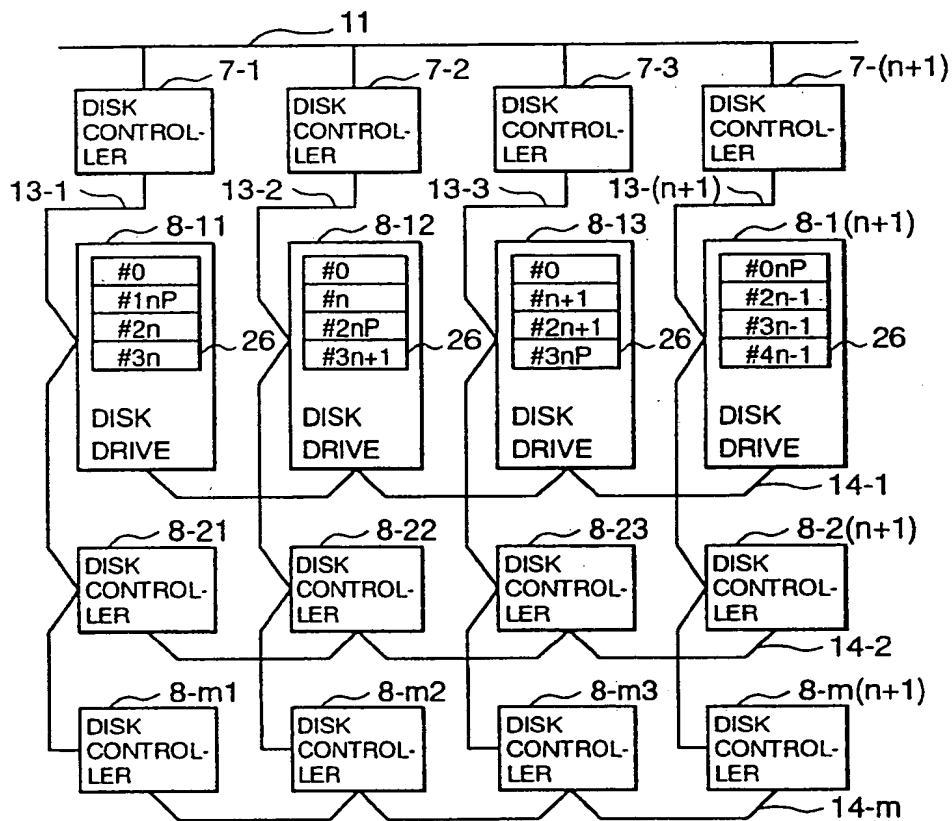


FIG. 6

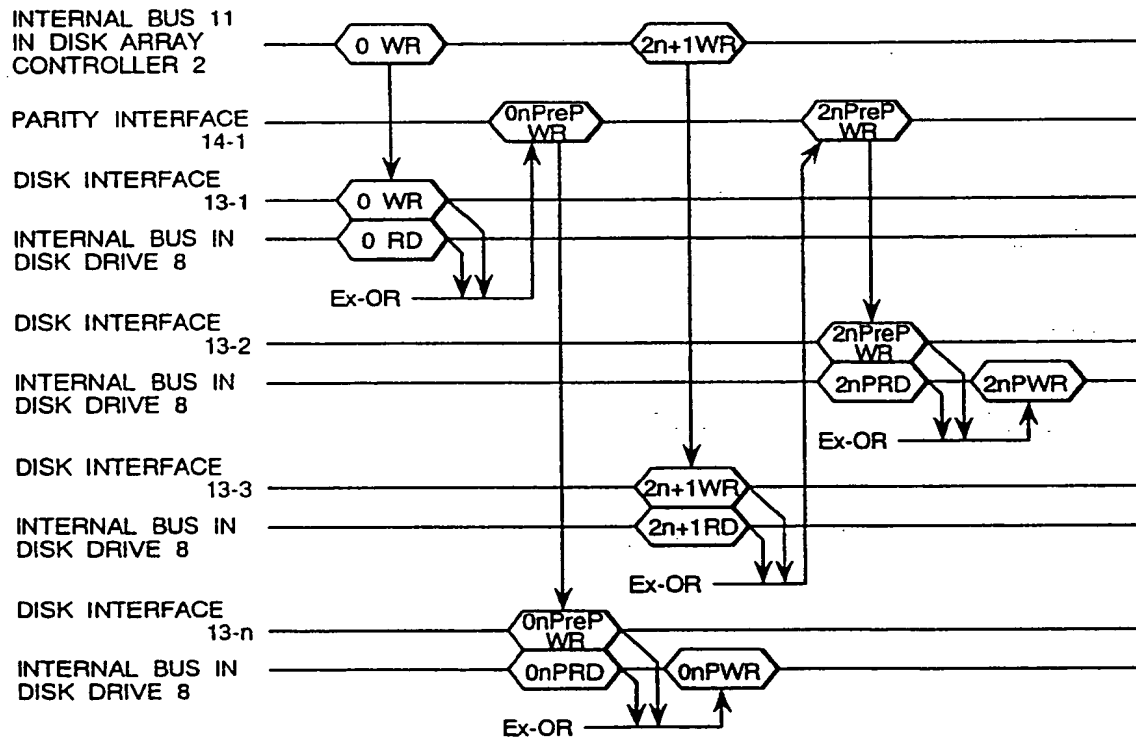
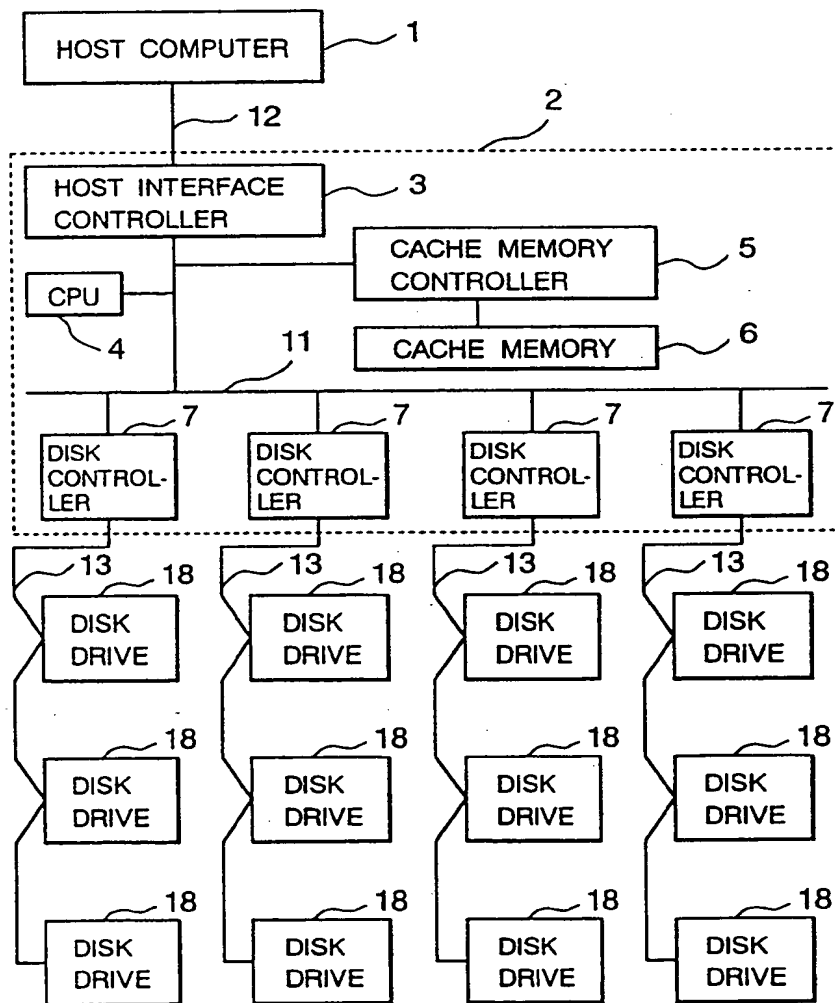
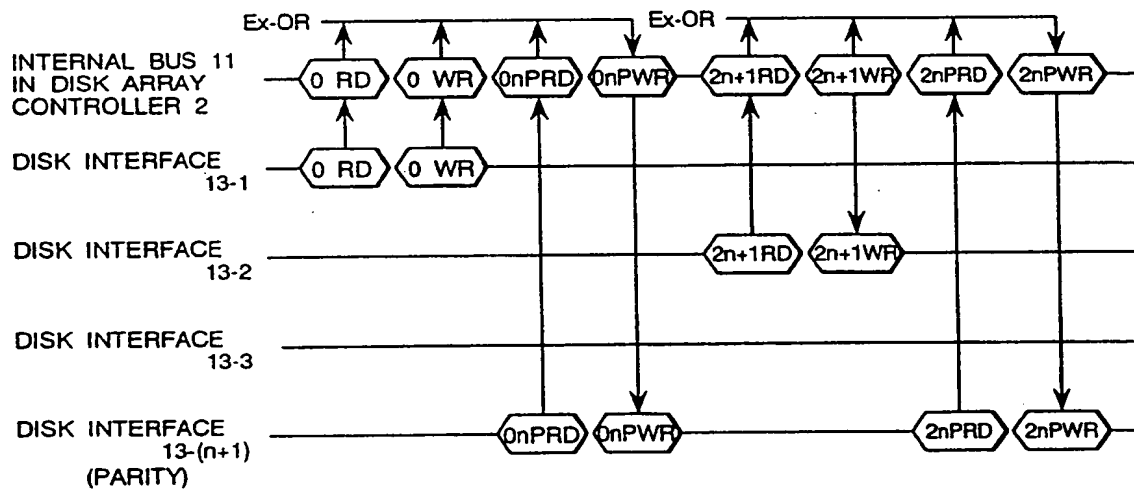


FIG. 7



**FIG. 8**





(19)



Europäisches Patentamt

European Patent Office

Office européen des brevets



(11)

EP 0 701 208 A3

(12)

## EUROPEAN PATENT APPLICATION

(88) Date of publication A3:

16.08.2001 Bulletin 2001/33

(51) Int Cl.7: G06F 11/10, G11B 20/18

(43) Date of publication A2:

13.03.1996 Bulletin 1996/11

(21) Application number: 95114142.3

(22) Date of filing: 08.09.1995

(84) Designated Contracting States:

DE FR GB

(30) Priority: 09.09.1994 JP 21559094

(71) Applicant: Hitachi, Ltd.

Chiyoda-ku, Tokyo (JP)

(72) Inventors:

- Kojima, Akira  
Odawara-shi, Kanagawa-ken 256 (JP)

- Ogino, Akihito  
Kanagawa-ken 259-01 (JP)

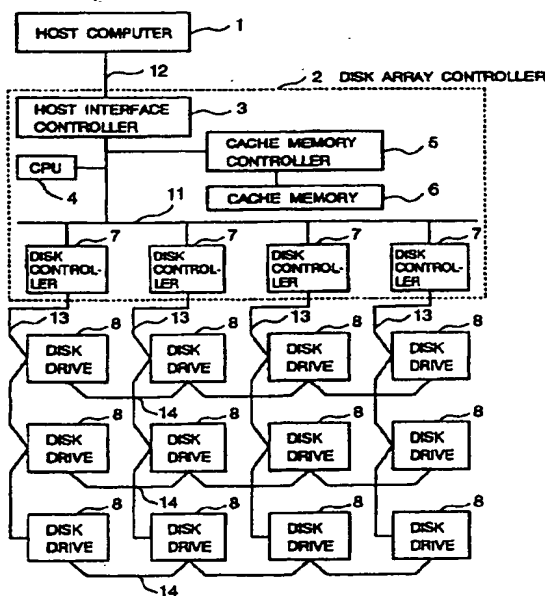
- Isono, Soichi  
Chigasaki-shi, Kanagawa-ken 253 (JP)

(74) Representative: Strehl Schübel-Hopf & Partner  
Maximilianstrasse 54  
80538 München (DE)

## (54) Disk array subsystem and data generation method therefor

(57) A disk drive has two channels of upper interfaces, and one channel is connected to a disk array controller, and the other channel is connected between a plurality of disk drives. A data disk drive reads the old data on the recording medium, calculates the exclusive OR of the old data and the corresponding data from the disk array controller, and transfers it to a parity disk as pseudoparity data from the other channel. The parity disk drive reads the old parity data on the recording medium, calculates the exclusive OR of the old parity data and the pseudo-parity data, and writes it as new parity data.

FIG. 1





European Patent  
Office

# EUROPEAN SEARCH REPORT

Application Number  
EP 95 11 4142

| DOCUMENTS CONSIDERED TO BE RELEVANT   |  |  |  |
|---|--|--|--|
| Category  | Citation of document with indication, where appropriate, of relevant passages  | Relevant to claim                                | CLASSIFICATION OF THE APPLICATION (Int.Cl.6) |
| X,D   | US 5 191 584 A (ANDERSON MICHAEL H)<br>2 March 1993 (1993-03-02)<br>* the whole document *   | 1-3,7  | G06F11/10<br>G11B20/18                       |
| A   | EP 0 550 853 A (MITSUBISHI ELECTRIC CORP)<br>14 July 1993 (1993-07-14)<br>* the whole document *   | 4  |  |
| A   | WO 93 18455 A (ARRAY TECHNOLOGY CORP)<br>16 September 1993 (1993-09-16)<br>* the whole document *  | 1-7  |  |
| P,X   | WO 94 22083 A (SCHAEFER UDO ;DRESCHER NORBERT (DE); SIEMENS NIXDORF INF SYST (DE))<br>29 September 1994 (1994-09-29)<br>* the whole document * | 1-3,7  |  |
|   |  |  | TECHNICAL FIELDS SEARCHED (Int.Cl.6)         |
|   |  |  | G06F<br>G11B                                 |
| The present search report has been drawn up for all claims  |  |  |  |
| Place of search<br>THE HAGUE  |  | Date of completion of the search<br>27 June 2001 | Examiner<br>Absalom, R                       |
| <p>CATEGORY OF CITED DOCUMENTS</p> <p>X : particularly relevant if taken alone<br/>Y : particularly relevant if combined with another document of the same category<br/>A : technological background<br/>O : non-written disclosure<br/>P : intermediate document</p> <p>T : theory or principle underlying the invention<br/>E : earlier patent document, but published on, or after the filing date<br/>D : document cited in the application<br/>L : document cited for other reasons<br/>&amp; : member of the same patent family, corresponding document</p> |  |  |  |

EPO FORM 1503 03 82 (P04C01)

**ANNEX TO THE EUROPEAN SEARCH REPORT  
ON EUROPEAN PATENT APPLICATION NO.**

EP 95 11 4142

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on  
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

27-06-2001

| Patent document<br>cited in search report | Publication<br>date | Patent family<br>member(s) | Publication<br>date |
|---|---------------------|----------------------------|---------------------|
| US 5191584 A                              | 02-03-1993          | CA 2098988 A               | 21-08-1992          |
|   |                     | DE 69227896 D              | 28-01-1999          |
|   |                     | DE 69227896 T              | 02-06-1999          |
|   |                     | EP 0572564 A               | 08-12-1993          |
|   |                     | WO 9215057 A               | 03-09-1992          |
| EP 0550853 A                              | 14-07-1993          | JP 2777301 B               | 16-07-1998          |
|   |                     | JP 5181611 A               | 23-07-1993          |
|   |                     | DE 69221279 D              | 04-09-1997          |
|   |                     | DE 69221279 T              | 05-03-1998          |
|   |                     | US 5544339 A               | 06-08-1996          |
| WO 9318455 A                              | 16-09-1993          | US 5398253 A               | 14-03-1995          |
| WO 9422083 A                              | 29-09-1994          | DE 4309148 A               | 29-09-1994          |

EPO FORM P0139

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

**THIS PAGE BLANK (USPTO)**